Improvement of k-Means Clustering Algorithm by GA

Khamis H Haroun[1], Wu Zhifeng[2]

[1,2]School of Information Technology Engineering, Tianjin University of Technology and Education, Tianjin, China 300222

Abstract: ***As known that K-means algorithm is the one of the common and popular technique for solving clustering problems. In fact, there are many k-means algorithms for solving the clustering problem such as Lloyd's k-means clustering algorithm, hierarchical k-means algorithm, also Grid based k-means algorithm etc. In the classical k-means algorithm the selected value of k must be confirmed first. So, the resulting clusters mainly depend on the selection of the initial centroids. It is not simple job to select the accurately value of k or to know exactly number of clusters for the given data set. So that in this paper propose new algorithm that called improvement of k-means clustering algorithm by GA that algorithm will be able to automatic find the best initial centers and appropriation of clusters according to the given data set.***

Keywords: Clustering, K-means, Cluster centroid, Genetic algorithm

1. Introduction

The term clustering can be defined as groping together an object that are similar to each another in the same cluster and also are dissimilar to those objects in another cluster. Or clustering is a just the process of classifying an object into different groups or it just the partition of a data set into subsets (clusters), so that the given data in each group (ideally) have nearly common trait often, due to some defined distance measured. Clustering is an unsupervised classification technique where there is no prior knowledge to that given data set is available. Cluster analysis is an important method in data mining field, hence it divides an unlabeled sample set into several sub-categories according to some sort of guidelines making similar samples grouped together and dissimilar samples divided into different classes as far as possible. Over the world the people have been create many applications related to the cluster analysis process in order to overcome the practical problems around us, in the really world example the researcher take the organized whether data to predicts the different whether conditions it is known that to understand the earth climate it require to find the patterns in the atmosphere and ocean so, that the cluster analysis is applied to find those required patterns in the atmosphere pressure of the polar regions and those areas of the ocean that have impact on the earth climate condition. Also, the cluster analysis process is mainly used in now days world advertisement purpose this is done after well organize and analysis the data based on customer's needs. Cluster analysis technique application helps the government to bring the social services in right place for the sake of reducing the distance travelling by the people to that place.

Genetic algorithm (GA) is a stochastic optimization technique in which by biological evolution. The genetic algorithm is a popular technique that has wide range of solving many kinds of problem, it mimic the process of biological evolution, and genetic algorithm is composed with the following operators, Selection operator is the operator in which the best solution has high probability to be selected for the reproduction, Crossover operator in this operator it produces the new off-spring (individual) by combination of the old ones and Mutation operator it involve the random changes to the individual.

2. K-means algorithm

2.1 Overview of the K-means algorithm

The K-means clustering algorithm is simply defined as a type of unsupervised learning that actually used when we have unlabeled data, (i.e. the data with no defined groups or categories). The main goal of the k-means clustering technique is to find the groups of the given data, with number groups always represented by the K variable. This algorithm words iteratively to allocate each of the given data point to the one of the K groups based on their features that provided before. The data points are clustered based on their similarities.
Consider the following set of data points below:

$$X = \{x_1, x_2, \dots, x_n\} \qquad x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ . \\ . \\ . \\ x_{id} \end{pmatrix}_{d\times 1} \quad (1)$$

The Set of the cluster is:
C = {c_1, c_2…c_k.} (2)
Also, the target is：
μk = mean of the cluster C_k
The squared error is:

$$J(c_k) := \sum_{x_i \in c_k} \|x_i - \mu_k\|^2 \quad (3)$$

The sum of squared errors is:

$$J(C) := \sum_{k=1}^{K} \sum_{x_i \in c_k} \|x_i - \mu_k\|^2 \quad (4)$$

The result of the K-means clustering algorithm contain the following:

Licensed Under Creative Commons Attribution CC BY

Inputs:
K: the number of clusters.
D: a data set containing n objects.

Output:
Set of k clusters.

Method:
Step 1: The randomly choose k objects from given data set as the initial cluster centers;
Step 2: Repeat
Step 3: Re assign every object to the cluster in which the object is the mostly closer or similar, this will be based on mean value of the elements in the cluster.
Step 4: Update the cluster means, this is done by calculate the mean value of the elements(objects) for each cluster.
Step 5: until no change

2.2 Description of K-means algorithm in a step wise

The algorithm contains of 4 main steps

1) Initialization
   This is a first step in which data set, number of clusters and the centroids that we defined for each cluster.
2) Classification
   In here the distance is computed for each and every centroid from a data point and for those data point with shortest distance from the centroid of a cluster could be assigned to that particular cluster.
3) Centroid Recalculation
   The recalculation is done as, if the data point is near to any of the centroid, then we have to compute the mean value and then change the old centroid value with the mean value and again compute (calculate) the distance with other data point and updated cluster centers.
4) The Convergence Condition Some convergence conditions are given as bellow:
   4.1 Stopping when the algorithm reaches a given or defined number of iterations
   4.2 Stopping when there is no again exchange of data points between the clusters
   4.3 Stopping when a targeted threshold value is achieved.
5) If incase all of the above conditions are not well satisfied, then go to step number 2 and the all steps repeat again, until when the given conditions are satisfied.

The existing k-means algorithm still has the main weakness of accuracy of convergence which is depend on approximation of the initial centroids [W.T.R Fernando. R. Wijewera and D.M.N.K Dasanayaka] and also the existing algorithm needs to specify the value of k in advance.

2.3 The main K-means weakness

The k-means algorithm is a good method to solve many clustering problems but still have main two problems which are:

1) The existing k-means algorithm is weak on finding the accuracy for the convergence.
2) Also, the k-means algorithm still has weakness to automatically approximate the number of cluster or the value of k, this means that in the existing k-means algorithm the user himself should specify the value of k in advance.

3. Genetic Algorithm

The genetic algorithm it originally developed by John Holland in (1975). Genetic algorithm is a search heuristic that imitate the process of evolution, it uses the concept of “Natural selection” and “Genetic inheritance” (Darwin 1859). The natural selection Darwin’s theory of evolution state that the only organism best adopted to their environment tend to survive and transmit their genetic characteristic on increasing number to succeeding generations while those less adopted tend to be eliminated. Biological our bodies are made up by trillions of cells, each cell has core structure called nuclear that contain our chromosomes. Also, the chromosomes are composed (made) up of tightly coiled strands of deoxyribonucleic acid (DNA). The genes are segments of DNA that determine specific traits, such as eye, hair or color. And we have more than 20,000 genes. A gene mutation is an alteration in your DNA, it can be transmitted from generation to generation (inherited) or acquired during your life time. The genetic algorithm is operating by maintain and manipulating the population of solutions with called chromosome [Sivananda &Deepa, 2008]. Every chromosome has its own fitness value which determine measure of the degree of the goodness of the solution encoded in it. We use the fitness value in order to guide us during the selection of the chromosomes, which are then used in generation of new offspring or candidate solutions through the crossover and mutation operators. The crossover operator it generates the new offspring by combine the two or more sections of the selected parents. For the mutation operator it is done by randomly selected a part of chromosome (gene) which is then altered, the three genetic algorithms operator’s selection, crossover and mutation will continue for a fixed number of generation or until a termination condition is meet. Genetic algorithm has wide range in application like in neural networks, machine learning, pattern recognition, image processing etc.

The flow chat is very important to show up the flow of activities of any process, so, here is the flow chat for the genetic algorithm technique.

Licensed Under Creative Commons Attribution CC BY

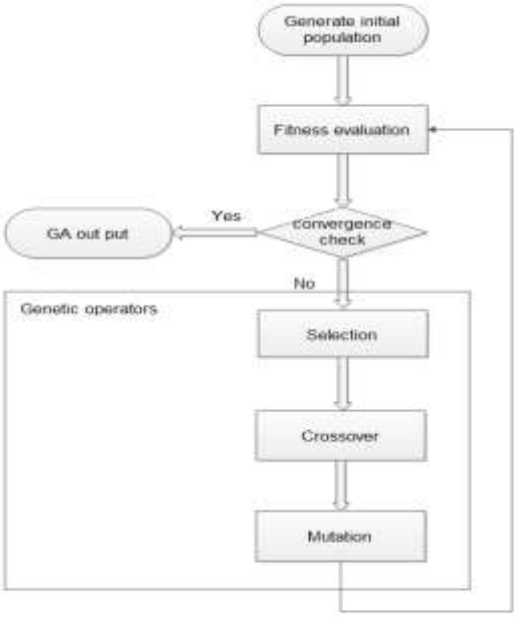


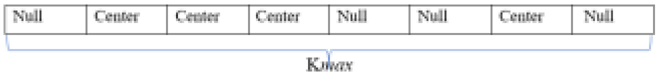
Figure 1: GA flow chart

4. The improved k-Means Clustering Algorithm by GA

Here let us see how the proposed method will be able to solve the problems of the existing algorithm, as known that the proposed method involves the genetic algorithm based on clustering technique to solve the existing problems, by using best approximation of the initial centroids and automatically appropriate the number of clusters or the value of K for the given data set, this means that no need to specify the value of K in advance. These two techniques of approximate the best initial centroids and automatic approximate the clusters number is possible under this genetic algorithm clustering technique.

4.1 Representation

This part explains how to encode the solution, let us assume k value to lie between the range of K*min* and K*max* in which the value of K*min* is 2, or otherwise is specified. The length of our string is K*max*, also the individual gene position in the chromosome will represent the actual center or a null value.



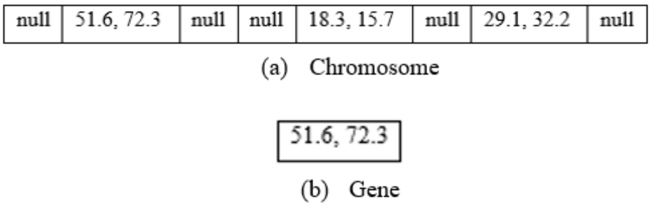
Null	Center	Center	Center	Null	Null	Center	Null

Figure 2: Representation of an individual

4.2 Population Initialization

The initialization of the population is done by randomly choose K*i* points from the given data set, the chosen points then are also randomly distributed to form a chromosome. This can be clearly explained by the example bellow.

Let us assume K*min* = 2 and the K*max* = 8. The random number K*i* = 3 for the chromosome *i*. This means that the chromosome will be encoded with 3 centers. Let the 3 clusters centers are (51.6, 72.3) (18.3, 15.7) (29.1, 32.2) that (randomly chosen from the given data set) and then we perform random distribution of those chosen points or centers in the chromosome, it may be look like this



null	51.6, 72.3	null	null	18.3, 15.7	null	29.1, 32.2	null

(a) Chromosome

51.6, 72.3

(b) Gene

Figure 3: Chromosome and Gene

4.3 Calculation of the fitness for individuals

The fitness can be accessed by involve two cases.

The first case is: there are clusters that will be formed or resulted during the encoding of the centers in the chromosomes. So that each point X$_i$, i =1, 2…, n, are assign the any one of the clusters C$_j$ which has the center Z$_j$ such that

$$\|X_i - Z_j\| < \|X_i - Z_p\| \text{ , P=1,2,...,K and j≠P} \tag{5}$$

Then when the clustering is done those cluster centers that have been encoded in the chromosome will be replaced by the points of the corresponding clusters. For the cluster C*i*, the new center will be calculated as

$$Z_i^* = \frac{1}{n_i}\sum X_i \qquad \text{i=1,2,...,K} \tag{6}$$

Consider the three points that have been chosen before during initialization of population that were (51.6, 72.3) (18.3, 15.7) (29.1, 32.2) with (51.6, 72.3) as their center, let the resulting cluster contains two more points, (50.0, 70.0) and (52.0, 74.0) besides itself i.e., (51.6, 72.3).

Therefore, the newly computed cluster center will become ((50.0+52.0+51.6)/3, (70.0+74.0+72.3)/3) = (51.2, 72.1). The new cluster center (51.2, 72.1) now replace the previous value of (51.6, 72.3).

$$\mu = \sum_{i=1}^{K} \mu_i, \tag{7}$$

$$\mu = \sum_{X_j \in C_i} \|X_j - Z_i\| \tag{8}$$

The fitness function is defined as:

$$F = \frac{1}{\mu} \tag{9}$$

This fitness function represents the sum of square of distances between the object and their centroids. So that when we maximize the fitness function this mean that we reduce or minimize the μ value. This is our aim to minimize the sum of squares of distances in clusters from their centroids.

Licensed Under Creative Commons Attribution CC BY

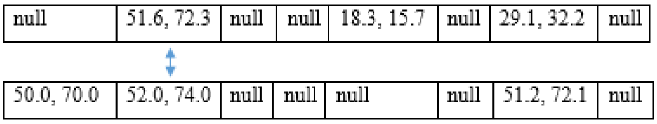
4.4 Genetic algorithm operators

a. Selection

It is commonly known that why we involve the selection operator in our algorithm this is because we need to reproduce more copies of the individual whose fitness values are higher. The selection operator plays important role to driving the search forward hence it will produce good solution within short period of time. In this algorithm conventional proportional selection will be applied on the population of strings. The chromosomes with higher fitness value is going to be selected for the production of the new offspring during the crossover operator. The rest of chromosome with lower fitness value will be removed or discarded, after the selection the best top three individuals from the rest will be stored for further use in mutation operator.

b. Crossover

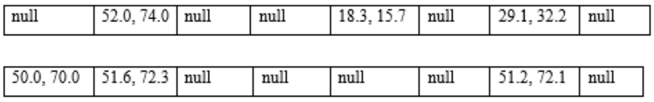
The crossover operator is one that is used to exchange the information among two different chromosomes in order to produce the two off springs in this paper the one-point crossover is applied with fixed probability value Pc = 0.8 as shown below with good example of the given chromosomes.



null	51.6, 72.3	null	null	18.3, 15.7	null	29.1, 32.2	null
50.0, 70.0	52.0, 74.0	null	null	null	null	51.2, 72.1	null

Figure 4: One-point crossover

Let the crossover position be 2nd positions for the given chromosome above. So that the resulting offspring will be look like



null	52.0, 74.0	null	null	18.3, 15.7	null	29.1, 32.2	null
50.0, 70.0	51.6, 72.3	null	null	null	null	51.2, 72.1	null

Figure 5: The resulted offspring

c. Mutation

The mutation operator will be done by random selection a part of chromosome (gene) within the chromosome of an individual, then after random select the mutation point to be changed we first check if the selected point (gene) is null or coordinates, if the gene is null it will be replaced by gene of the coordinates from any of the best top three individual stored during the selection operator. Or if the selected gene contains coordinate (not null) it will be replaced by null gene from any of the best top three individuals.

5. Experiments and Results

In this part or section will contain some activities that have been carried out to obtain the output results, it will include the name and description of the data set used during the during the experimental analysis process and some parameters used in proposed algorithm.

5.1 Data set description

As we know that the data set are very important resource in the research during the experimental analysis process in order to get the results on what kind of problem you try to solve. So, that in my research there are three-artificial data sets that involve in the experimental analysis, the name for the datasets are Circular_4_3, Circular_5_2 and Circular_6_2. Those three artificial data sets were generated using the uniform distribution, also, the name of each data set will imply the structure of the classes and concatenated with the number of clusters that are actually presented in the data set itself with number of dimensions. For example, let us take the data set called Circular_5_2 in this data the clusters appear to be circular in nature, so, there are 5 clusters in this data set and also the dimension of the cluster is 2.

Also, the experiment involves one data set from the UC Irvine repository which is called iris data set, the iris data set is composed of 150 as number of objects(element) in it, with 4 attributes namely as Width Sepal, Petal Length, Sepal Length Width Petal. The iris data set has got three different classes which are Iris virginica, iris Setos, and Iris versicolor, the actual number of clusters in the iris data set is 3.

The table below summarize the details explanation on those three-artificial and iris data set used as follow below

Table 1: Shows summary of the data set used

S.No	Data set name	Number of objects in data set	Number of clusters	Dimension
1	Circular_4_3	1200	4	3
2	Circular_5_2	500	5	2
3	Circular_6_2	600	6	2
4	Iris	150	3	4

Note

It is known that before starting clustering the dataset the first thing which is very important to do is pre-processing the data this is because, the data are various type with different attributes which have different value for different types, range, so, that we need to do pre-processing by convert the categorical attribute to binary attributes, for example, an attribute indicating the sex (male or female) here 1 can be used to represent male and 0 to represent female.

5.2 Results

Before display the results of the proposed method it's better to give definition on the type of parameter used during the experimental analysis. And brief explanation on what kind of data set used as explained above, also the obtained results of the proposed method will be compared with classical k-means algorithm or the k-mean algorithm without genetic algorithm. The implementation of the proposed method is done under the following parameters.
The maximum iteration is: 70.
The population size is: 100.
Crossover percentage is: 0.8
Mutation rate is: 0.02

Licensed Under Creative Commons Attribution CC BY

The table below include the result that observed by the new proposed method and then compared with the result obtained by the algorithm without GA.

Table 2: Result of the proposed method

Data set Number	Objects in Data set	Number of clusters detected by GA	Iteration of K-means without GA	Iteration of K-means with GA
1	1200	4	70	30
2	500	5	97	65
3	600	6	102	70
4	150	2	39	25

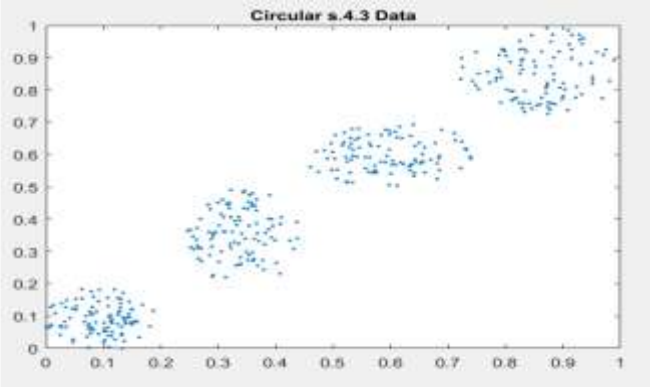


Figure 6: Circular_4_3 data set distribution.

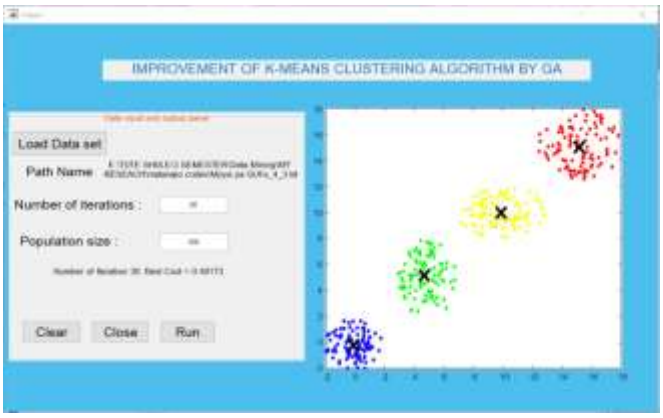


Figure 7: The better clustering done by proposed method for the Circular_4_3 data set.

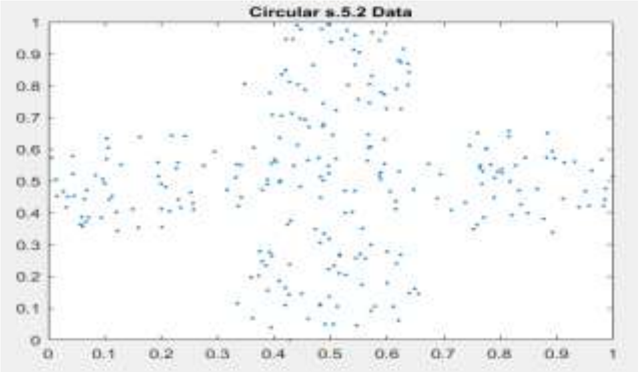


Figure 8: Circular_5_2 data set distribution.

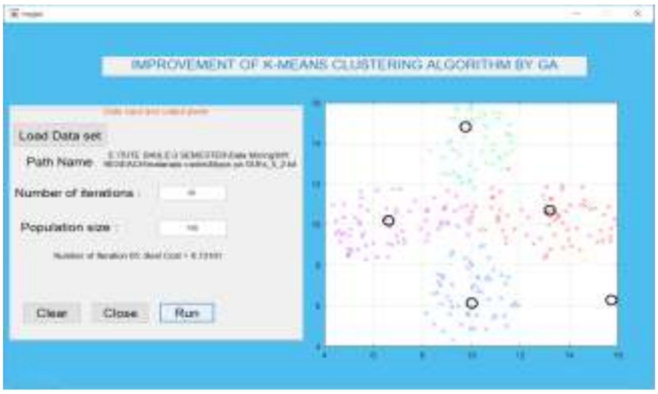


Figure 9: The better clustering done by proposed method for the Circular_5_2 data set.

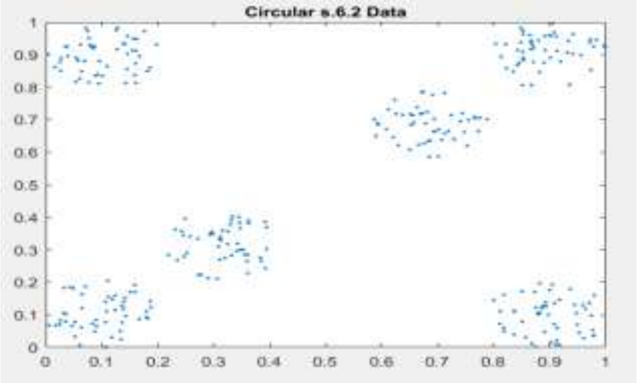


Figure 10: Circular_6_2 data set distribution.

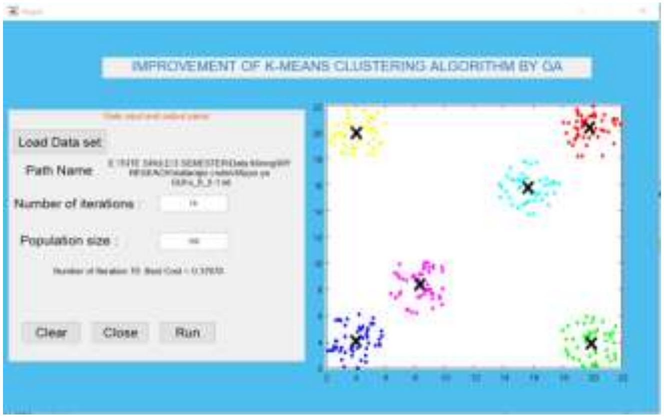


Figure 11: The better clustering done by proposed method for the Circular_6_2 data set.

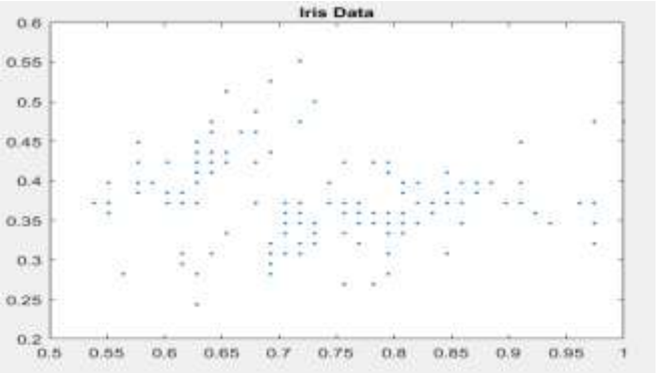
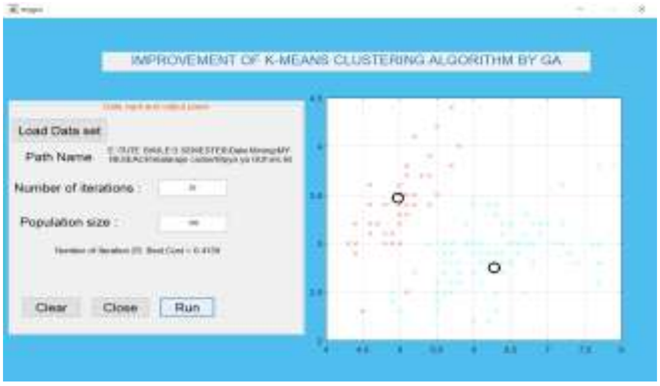


Figure 12: Iris data set distribution.

Licensed Under Creative Commons Attribution CC BY



IMPROVEMENT OF K-MEANS CLUSTERING ALGORITHM BY GA

Load Data set

Path Name

Number of iterations :

Population size :

Clear Close Run

Figure 13: The better clustering done by proposed method for the Iris data set.

Based on our table 2 and figures 7, 9, 11, 13 above, it can be seen clearly and judge the performance of the proposed method, which involves the genetic algorithm, the method gave us better clustering result compare to the that clustering method without genetic algorithm.

1. Conclusion

In this proposed method it explains how to solve the problem of accuracy of convergence in k-means clustering algorithm by using the genetic algorithm technique. The method introduces how to find initial centroid of the cluster for the given data set using GA which is very power full technique that is mainly used in many clustering algorithms. But for the best result the most modified genetic algorithm operator should be studded and technical implemented for further improving of this k-means clustering algorithm.

References

[1] Jiawei Han, Michelin Kamber, Jian Pei.”DATA MINING Concepts and Techniques” third addition, (ISBN 978-0-12-381479-1).

[2] http://www.slideshare.net/parryprabhu/k-meanclustering-algorithm?/On2017/05/01.

[3] http://www.slideshare.net/parryprabhu/k-meanclustering-algorithm.

[4] Anil K. Jain. “Data Clustering: 50 Years Beyond K-Means[1]” Department of Computer Science & Engineering, Michigan State University East Lansing, Michigan 48824 USA.

[5] V.Ilango, Dr.R.Subramanian, Dr.V.Vasudevan “Cluster Analysis Research Design model, Problems, issues, challenges, trends and tools” International Journal on Computer Science and Engineering (IJCSE).

[6] Petra Kudová, “Clustering Genetic Algorithm” Department of Theoretical Computer Science Institute of Computer Science Academy of Sciences of the Czech Republic ETID 2007.

[7] Rouhollah Maghsoudi[1], Arash Ghorbannia Delavar[2], Somayye Hoseyny[3], Rahmatollah Asgari[4], Yaghub Heidari[5] “Representing the New Model for Improving K-Means Clustering Algorithm based on Genetic Algorithm” Department of Computer, Nour Branch, Islamic Azad University, Nour, Iran, The Journal of Mathematics and Computer Science Vol .2 No.2 (2011) 329-336.

[8] Tsai-Yang Jea, “Basic concepts of Data Mining, Clustering and Genetic Algorithms” Department of Computer Science and Engineering SUNY at Buffalo.

[9] Zheyun Feng, “Data Clustering using Genetic Algorithms” Department of Computer Science and Engineering, Michigan State University

[10] M.Anusha, J.G.R.Sathiaseelan, “An Enhanced K-Means Genetic Algorithms for Optimal Clustering” Department of Computer Science, Bishop Heber College, Trichy-17, Tamilnadu, INDIA.

[11] Ke Chen, “K-means clustering” The University of Manchester, COMP24111 Machine Learning

[12] K. A. Abdul Nazeer, M. P. Sebastian, “Improving the Accuracy and Efficiency of the k-means Clustering Algorithm” Proceedings of the World Congress on Engineering 2009 Vol I WCE 2009, July 1 - 3, 2009, London, U.K.

[13] Khaled Alsabti, Syracuse University. Sanjay Ranka University of Florida. Vineet Singh Hitachi America, Ltd. “An Efficient K-Means Clustering Algorithm”.

[14] Wojciech Kwedlo and Piotr Iwanowicz “Using Genetic Algorithm for Selection of Initial Cluster Centers for the *K*-Means Method” Faculty of Computer Science, Bia_lystok University of Technology, Wiejska 45a, 15-351 Bia_lystok, Poland

[15] K.Arun Prabha, R.Saranya, “Refinement of K-Means Clustering Using Genetic Algorithm” Journal of Computer Applications (JCA) ISSN: 0974-1925, Volume IV, Issue 2, 2011.

[16] G.Kiran Kumar, T.Bala Chary, Department of CSE, MLRIT, Hyderabad A.P, India. P.Premchand Department of CSE, University College of Engineering, Hyderabad, A.P, India.” A New and Efficient K-Means Clustering Algorithm” International Journal of Advanced Research in Computer Science and Software Engineering. ISSN: 2277 128X, Volume 3, Issue 11, November 2013.

[17] Anand M. Baswade[1], Prakash S. Nalwade[2],[1]M.Tech, Student of CSE Department, SGGSIE&T, Nanded, India [2]Associate Professor, SGGSIE&T, Nanded, India,” Selection of Initial Centroids for k-Means Algorithm” International Journal of Computer Science and Mobile Computing ISSN 2320–088X IJCSMC, Vol. 2, Issue. 7, July 2013, pg.161 – 164.

[18] Hu Shu-chiung, Genetic algorithm-based clustering technique, Ujjwal Maulik, Sanghamitra Bandyopadhyay, 2004.

Author Profile

Khamis Hamad Haroun received the Bachelor Degree in Information System and Network Engineering from St. Joseph University in Tanzania in 2013. Currently pursuing the M.E in Applied Computer Technology from Tianjin University of Technology and Education. Mr. Khamis was working as tutorial assistant at St. Joseph University in Tanzania from February 2014 – August 2014. He based on computer vision, data mining, machine learning as research areas.

Licensed Under Creative Commons Attribution CC BY

Wu Zhifeng: Professor at Tianjin University of technology and Education. His research interests include the evolution computation, machine learning, knowledge discovery and data mining. He obtained his Ph.D. at Beijing Jiaotong University.

Licensed Under Creative Commons Attribution CC BY